

PadhAI: Backpropagation - the full version

One Fourth Labs

Computing derivatives w.r.t all weights in any layer

1. Let's take a simple example of a $W_k \in \mathbb{R}^{3 \times 3}$ and see what each entry looks like

a.

$$\nabla_{W_k} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial W_{k11}} & \frac{\partial L(\theta)}{\partial W_{k12}} & \frac{\partial L(\theta)}{\partial W_{k13}} \\ \frac{\partial L(\theta)}{\partial W_{k21}} & \frac{\partial L(\theta)}{\partial W_{k22}} & \frac{\partial L(\theta)}{\partial W_{k23}} \\ \frac{\partial L(\theta)}{\partial W_{k31}} & \frac{\partial L(\theta)}{\partial W_{k32}} & \frac{\partial L(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \nabla_{W_{kij}} L(\theta) = \frac{\partial L(\theta)}{\partial a_{ki}}$$

$$\nabla_{W_k} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial L(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial L(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial L(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial L(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} = \nabla_{a_k} L(\theta) \cdot h_{k-1}^T$$

2. Thus we can update all the weights in one go using $W_k = W_k - \eta(\nabla_{a_k} L(\theta) \cdot h_{k-1}^T)$ a

3. $\nabla_{W_k} L(\theta) = \nabla_{a_k} L(\theta) \cdot h_{k-1}^T$

4. Finally coming to the biases, $a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$

5. $\frac{\partial L(\theta)}{\partial b_{ki}} = \frac{\partial L(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}} = \frac{\partial L(\theta)}{\partial a_{ki}}$

6. We can now write the gradient w.r.t the vector b_k

a.

$$\nabla_{b_k} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{k1}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{kn}} \end{bmatrix} = \nabla_{b_a} L(\theta)$$

7. Thus, we can update all biases using $b_k = b_k - \eta(\nabla_{b_k} L(\theta))$ a

8. $\nabla_{b_k} L(\theta) = \nabla_{a_k} L(\theta)$